ESCHERICHIA COLI AND SALMONELLA TYPHIMURIUM

CELLULAR AND MOLECULAR BIOLOGY

VOLUME 2

Editor in Chief

Frederick C. Neidhardt
Department of Microbiology and Immunology
University of Michigan, Ann Arbor, Michigan

Editors

John L. Ingraham
Department of Bacteriology
University of California
Davis, California

K. Brooks Low
Radiobiology Laboratories
Yale University
New Haven, Connecticut

Boris Magasanik
Biology Department
Massachusetts Institute of Technology
Cambridge, Massachusetts

Moselio Schaechter
Department of Molecular Biology and Microbiology
Tufts University School of Medicine
Boston, Massachusetts

H. Edwin Umbarger
Department of Biological Sciences
Purdue University
West Lafayette, Indiana

AMERICAN SOCIETY FOR MICROBIOLOGY
Washington, D.C.

56. Genome Organization

MONICA RILEY[1] AND STEVEN KRAWIEC[2]

Department of Biochemistry, State University of New York at Stony Brook, Stony Brook, New York 11790,[1] and Department of Biology, Lehigh University, Bethlehem, Pennsylvania 18015[2]

INTRODUCTION

The chromosomal DNAs of *Escherichia coli* and *Salmonella typhimurium* share many properties. A single, continuous, negatively supercoiled molecule is folded into a nucleoid structure that bends at specific sequences and forms nucleosomelike structures by interaction with histonelike protein molecules (for minireviews, see references 27, 69, and 70).

Extensive electrophoretic characterizations of allozymes from more than 1,600 isolates of *E. coli* have revealed "strong multilocus associations" (86, 87); in other words, there are only a few clonal types of *E. coli* bacteria, and there seems to be little genetic exchange among these enteric organisms in their natural envi-

ronments. Despite the stability of the genomes of these contemporary organisms, their chromosomes, like those of other organisms, have changed during evolutionary time. In this chapter, detailed comparisons of the sequence of loci in *E. coli* and *S. typhimurium* are presented, and the means by which chromosomal material in these organisms can be acquired and rearranged are discussed.

CONSERVED FEATURES

Size and Complexity

Physical measurements

The near equivalence of *E. coli* and *S. typhimurium* genome sizes is evident when they are compared with the range of sizes that occur among procaryotes (6, 9). The smallest procaryotic genome, 0.4×10^9 daltons, occurs in *Chlamydia trachomatis*, and the largest, 8.6×10^9 daltons, occurs in *Calothrix* spp. (45). The genomes of *E. coli* and *S. typhimurium* form a cluster within this spectrum. By relating the genetic map distances with physical measurements of size of the corresponding DNA, the conversion factor of 45 to 46 kilobases (kb) per map unit has been obtained for two different, extensive regions of the *E. coli* K-12 chromosome. If these regions are representative of the whole, and if all map units are comparable, this conversion factor yields a value of 3.0×10^9 daltons for the 100-map-unit chromosome of *E. coli* K-12 (11, 41).

Renaturation

Approaching the characterization of bacterial DNA from another direction, the degree of complexity of the DNA has been estimated from second-order renaturation kinetics (14, 117). Values were obtained that ranged from approximately 3.8×10^6 base pairs (bp) in unique sequences for *E. coli* C (13, 37) through a value of 4.7×10^6 bp for an isolate of *S. typhimurium*. Extensive data indicate that the overall complexities of *Salmonella* strains are generally 10 to 20% greater than that of *E. coli* K-12 (24). Despite this level of variability, some of the values obtained for some of the strains were nearly identical for the two bacteria. For example, the complexity of *E. coli* K-12 was determined to be 4.0×10^6 bp, and that of *S. typhimurium* LT7 was 4.2×10^6 bp. It may be noted that these complexities refer to the entirety of the complements of unique sequences of DNA in the cell, the sum of both plasmid and chromosomal elements. The presence of a plasmid in one organism and the absence of an equivalent structure in a second organism could contribute to greater or lesser levels of perceived similarity in complexity.

Besides unique sequences, part of the genomic DNA is in the form of repeated sequences. Experiments on the fast-renaturing fraction of the DNA have revealed that 0.5 to 2% of the sequences in the *E. coli* genome may be "foldback" DNA, or closely linked reiterated sequences (21, 61, 71).

In sum, the results indicate that there is some variation in the measure complexities of genomes in any one species but that the sizes of genomes in strains representing both *E. coli* and *S. typhimurium* are approximately equivalent.

Base Composition

Uniformity

The base compositions of bacterial DNAs differ one from another, but the chromosomal DNA of any one bacterial strain is relatively uniform in composition throughout its genome. The width of the distribution curve for DNA fragments that have been centrifuged to equilibrium in CsCl gradients is a function of the size distribution of the fragments and of the extent of heterogeneity in base composition. Judging by the width of the profiles of equilibrium bands of bacterial DNAs, the guanine-plus-cytosine (GC) contents cluster closely about distinct mean values (95). Because unusually broad CsCl DNA profiles have not been observed, extremes of base composition must be confined to nucleotide sequences that are short relative to the lengths of shear fragments of chromosomal DNA ($>10^6$ bp), and the majority of DNA shear fragments must possess GC contents that are close to the chromosomal average.

Most *S. typhimurium* strains are 1 to 4% higher in GC content than are most *E. coli* strains (103). When the GC contents of both *E. coli* K-12 and *S. typhimurium* LT2 were measured in the same series of experiments by determining either the thermal denaturation temperature or the buoyant density in CsCl, the *E. coli* K-12 DNA was found to be 1.5 to 2.0% lower in GC content than was *S. typhimurium* LT2 DNA (7, 77, 99).

Bias in third position of codon

There appears to be a mechanism that conserves the GC content of the genome of a given bacterial strain within a relatively narrow range. The explanation of how the GC content of bacterial DNA is established, maintained, or modified is not yet at hand, but a part of the mechanism seems to lie in the selection of synonymous codons having suitable GC content. Comparison of the nucleotide sequences of corresponding genes in *E. coli* and *S. typhimurium*, such as genes of the *trp* operon, supports this view. In both of these bacteria, the *trp* structural genes are higher in GC content than the average value for the respective genomes. Each of the *S. typhimurium* genes has a higher GC content relative to each corresponding *E. coli* gene (Table 1). Almost half of all nucleotide differences between the *E. coli* and *S. typhimurium trp* genes occur in the third positions of codons. Most of the nucleotide differences in the genes of *E. coli* versus the genes of *S. typhimurium* generate synonymous codons. Selective codon usage clearly can be a tool for establishing or adjusting GC content.

When one examines GC content at higher levels of organization, e.g., contrasting structural genes with regulatory regions, heterogeneity in base composition is revealed. On the one hand, some regulatory sequences are AT rich (e.g., control sequences for the *lpp* and *ompF* genes [56, 79] and noncoding sequences upstream of *leuABCD* in both *E. coli* and *S. typhimurium* [43]). On the other hand, among individual structural genes some are higher in GC content and

TABLE 1 GC content of *trp* structural genes of *E. coli* K-12 and *S. typhimurium* LT2

Genes (reference)	GC content (%)	
	E. coli K-12	*S. typhimurium* LT2
trpA codons (80)		
All positions	54	57
Third positions	56	63
trpB codons (23)		
All positions	54	57
Third positions	60	64
trpG(D) (81)		
All positions	55	57
Third positions	61	62
trpE (120)		
All positions	54	58
Third positions	55	62

some are lower in GC content than the chromosomal average. These departures from the average affect genetic entities that are small ($<10^6$ daltons) relative to the entire chromosome, suggesting that the mechanisms that determine GC content of the genome are able to sense localized heterogeneity of a gene or an intercistronic region and to determine the degree of compensation that is needed to counterbalance any small aberrant regions. The mechanism that a cell possesses for sensing GC content and for invoking compensatory adjustments has yet to be discerned.

Gene Order

Coincidence of loci

Genetic maps of related organisms tend to be similar (62). Sanderson (96) noted that there are many similarities in the genetic maps of *E. coli* and *S. typhimurium*. More genes have been mapped in both organisms since that time, and it remains true that the two maps are congruent. A previous exercise in map comparison (94) has been brought up to date here by using the most recently compiled map information (5, 97) (see Fig. 1) There are some exceptions with respect to gene order and distances between markers that will be discussed below, but, in general terms, a remarkable degree of congruence has survived in the genomes of two organisms which are believed to have diverged from one another many millions of years ago, and the DNAs of which have replicated independently of each other many millions, perhaps billions, of times.

Since *E. coli* and *S. typhimurium* are able to engage in sexual recombination with each other only at very low frequency, it is doubtful that the need to preserve opportunities for homologous recombination over long linkage distances has served as a stabilizing effect on gene order. Instead, the conservation of gene order implies that there is a functional correlate of this gene order that confers a growth or survival advantage on the organism. Even in an organism as far removed from the enteric bacteria as *Bacillus subtilis*, the arrangement and function of a set of genes, those near the origin of replication, exhibit important similarities with the corresponding genes in *E. coli* (88), suggesting that function and gene location are not entirely independent for this set of genes.

Chromosomal rearrangements

Whatever the (unknown) advantage of a particular gene order may be, it is not so powerful that it prevents survival of some *E. coli* or *S. typhimurium* variants that have undergone rearrangement of parts of their genomes. In fact, a large inversion of approximately 15% of the genome has occurred in the evolutionary past in either *E. coli* or *S. typhimurium* with no known ill effect (19). There are other instances in which variants have arisen spontaneously, having sustained rearrangements within their genomes. Inversion of 17% of the genome in *E. coli* W2637 and its descendant, W3110, arose spontaneously as a result of recombination between two nearly identical *rrn* loci in the *E. coli* genome (50). Another large inversion was discovered in the *lac* region of *E. coli* SY99, apparently produced by recombination between two copies of IS3 that flank the inverted segment (98). In addition to inversions, duplications also arise spontaneously in laboratory strains. In the course of a study of the *recB* and *recC* genes, a duplication in the *thyA-argA* region was discovered in *E. coli* KL399 (29). These examples of spontaneous changes in map locations of chromosome segments and of individual genes show that even though the order of most genes in the chromosomes of *E. coli* and *S. typhimurium* has been preserved, change is not excluded altogether. Rearrangements seem to be discouraged in the sense that they are rare events, yet they are not completely forbidden.

In the laboratory, selections have been designed with an eye to detecting rearrangement mutations (e.g., references 51, 100). Inversion mutants have been isolated, and characterization has shown that there are constraints on the location of the endpoints of inversions that seem to relate in some way to the spatial relationship between the origin and the terminus of replication. When selection for a rearrangement involving the *his* operon was exerted, inversion mutants were obtained, although at much lower frequency than duplications. Endpoints of the inversions were not located at random, but were such as to bring the terminus of replication closer to the origin (100).

Integrity of the terminus of replication in the chromosome has been shown to be important for the health of the cell. Deletion of 340 kb of DNA from the terminus region of the chromosome interfered with control of the travel of replication forks and caused poor growth and reduced viability (44). When a plasmid was integrated into the chromosome near the terminus of replication, some *E. coli* strains suffered physiological damage and became sensitive to growth on rich medium (75). Revertants resistant to rich medium were isolated, and some of these were found to have undergone major chromosomal rearrangements: inversions of approximately half of the chromosome along the axis of the origin and terminus of replication. The phenotypic effects on the rate of growth in rich medium seem to be the consequence of altered velocities of replication forks (75). Thus, at least some chromosomal rearrangements can be shown to confer a growth disadvantage on the orga-

nism, providing a partial explanation for the tendency to conserve global gene order

Secondary Structure

Secondary structures in the transcripts of structural genes are important features of mechanisms for initiating, terminating, and regulating gene expression. Comparison of regulatory sequences of analogous genes in *E. coli* and *S. typhimurium* has shown that, in regulatory regions, secondary structures are well conserved even when base sequence has undergone change. Despite nucleotide sequence variation, the alternate stem-and-loop structures of the attenuation regions of the *trp* operon were conserved in a group of enteric bacteria, among them *E. coli* and *S. typhimurium* (119). Similarly, sequences at the 3' end of the *trp* operon in *E. coli* and *S. typhimurium* are not highly conserved in base sequence, but the stem-and-loop structure for the transcript terminator has been retained (81). The sequences and secondary structures of the leader sequences of the *ilvGEDA* operons of *E. coli* and *S. typhimurium* are highly conserved (42). Also, the inverted repeat sequences that provide the potential for stem-and-loop structures in transcripts of intercistronic regions in the *his* operon and in the *malB* region have been highly conserved (36, 46).

PATTERNS OF REPEATED SEQUENCES

Recurrent sequences, which occur at different organizational levels, appear to be significant in influencing the structure of the genome. There are gene duplications, short duplicated segments within genes, large (>1-kb) directly repeated or inverted sequences, and small (<50-bp) repeated sequences that are located both within and between coding sequences.

Gene Duplication

Multicopy genes

rrn loci (which commonly have the sequence promoter$_1$-promoter$_2$-16S rRNA-tRNA-23S rRNA-5S rRNA-tRNA-terminator$_1$-terminator$_2$) occur at several sites on the chromosomes. Both *E. coli* and *S. typhimurium* have seven *rrn* loci which are present at equivalent positions in the two genomes (66). Furthermore, the left-to-right orientations of the loci in the chromosomes in these two enteric bacteria are the same. Whereas the overall structure of these loci is highly conserved, differences based on the identity of tRNA in the spacer region have been recognized. In particular, four loci in both *E. coli* and *S. typhimurium* contain $tRNA_2^{Glu}$, and three contain $tRNA_{1B}^{Ala}$ and $tRNA^{Ile}$. Interspecific comparisons reveal that *rrn* loci at equivalent positions in the genome do not necessarily contain the same tRNA. In *E. coli*, *rrnD* contains $tRNA_{1B}^{Ala}$; in *S. typhimurium*, the *rrnD* locus contains $tRNA_2^{Glu}$ (66). The reciprocal relationship exists in *rrnB*; in *E. coli* this locus contains $tRNA_2^{Glu}$, whereas in *S. typhimurium* $tRNA_{1B}^{Ala}$ is present (66). Because of the high degree of homology among the *rrn* loci, one might expect frequent exchanges of segments among *rrn* loci. In fact, rearrangements are rare, and the organizations of these loci are remarkably stable.

Duplicated genes

Sequence similarities (homologies) are evident in the nucleotide sequences of many pairs or groups of structural genes or in the amino acid sequences of their protein products. Such genes are believed to be related by duplication of an ancestral gene followed by divergence of sequence and function. The most extensively documented examples pertain to biosynthetic operons and to the sensory transducers that modulate responses to attractants and repellants in the environment. Pairs and groupings of structural genes which have extensive homologies are listed in Table 2. Selected examples will be used here to illustrate some general features of duplicated genes.

Ornithine and aspartate carbamoyltransferases. The genes for the enzymes ornithine carbamoyltransferase and aspartate carbamoyltransferase in *E. coli* exhibit levels of homology that are to be expected either between recently duplicated genes or between ancestrally related genes (53, 114). Ornithine carbamoyltransferase is specified by two loci in *E. coli* K-12, *argI* located at map position 97 and *argF* at position 7 (5). A total of 78.1% of the nucleotides in the two genes are identical, and 86% of the 333 amino acids of the enzymes are identical. It seems likely that these two genes arose recently by duplication followed by divergence and transposition.

The *argI* locus for ornithine carbamoyltransferase is linked to the *pyrB* locus, which specifies aspartate carbamoyltransferase. This latter enzyme contains 310 amino acids. By introducing gaps, the codons in *pyrB* can be aligned with the nucleotide sequences for ornithine carbamoyltransferase. Such a comparison reveals that *pyrB* is 35 to 40% homologous to both *argI* and *argF* (53, 114). Among these three genes, *pyrB* has diverged from *argF* and *argI* more than the latter two have from each other. The pattern of divergence suggests that the duplication event that separated the ornithine carbamoyltransferase and aspartate carbamoyltransferase genes preceded the duplication that generated the two genes for ornithine carbamoyltransferase.

Chemoreceptors. In *E. coli*, four loci are known which specify chemoreceptors (*tar*, *tap*, *tsr*, and *trg*). Just as for the *arg* and *pyr* genes discussed above, the loci that are genetically linked (*tar* and *tap*) are less related to one another than either is to the unlinked *tsr*

TABLE 2. Sets of closely related genes presumed to have arisen by duplication of an ancestral gene

Genes	Reference(s)
argF, argI, pyrB	53, 114
aroF, aroG	25, 102
hisJ, argT	47
ilvBN, ilvGM, ilvIH	35, 106, 116
lysC, metL, thrA	18, 34
metB, metC	8
ompC, ompF, phoE	56, 78, 89
tar, tap, tsr, trg	10, 12, 63
trp(G)D, pabA	52, 60
trpE, pabB	39, 82
tufA, tufB, fus	2, 121, 123
sucB, aceF	105, 107
xap, desD	17

locus (63, 115). (The fourth locus, *trg*, is even less related [10].) Comparison of the chemoreceptor sequences reveals that levels of homology are not uniform throughout the length of a locus. The protein products have been divided into six domains; some are common to the various proteins (e.g., a methylation region), and others are distinct in each protein (e.g., attractant recognition sites). The distinctive portions are concentrated in the amino end of the protein, whereas the common regions tend to be at the carboxyl end. As a consequence, levels of amino acid sequence similarity are low (10 to 60%) at the amino end and high (60 to 100%) at the carboxyl end.

It should be noted that common function in a protein does not necessarily signify nucleotide homologies in the corresponding genes. Among the various sensory-transducing proteins, there is little homology among the membrane-spanning segments. The low level of identity may reflect a low level of specificity required for the function. The membrane-spanning section just precedes the region of high homology. Presumably, some functionally important segments of ancestral genes have been conserved while other portions with more flexible requirements have diverged.

Cystathionine-γ-synthase and β-cystathionase. The unlinked genes *metB* and *metC* encode adjacent enzymes in the pathway of synthesis of methionine. Nucleotide sequences have been determined, amino acid sequences have been deduced, and sequences have been compared (8). Sequence similarities are evident, distributed throughout both the genes and the proteins, suggesting that sequential enzymes in a metabolic pathway could have arisen by gene duplication and divergence.

Aspartokinase-homoserine dehydrogenase. In *E. coli*, the *thrA* and *metL* loci specify, respectively, the bifunctional isozymes aspartokinase I-homoserine dehydrogenase I and aspartokinase II-homoserine dehydrogenase II (34). The tetrameric enzymes have three domains; one at either end of the polypeptide chain for each of the two catalytic activities and a central domain concerned with subunit interaction. The two bifunctional enzymes have similarities of structure and sequence that suggest strongly that the two genes arose by duplication of an ancestral bifunctional gene (34).

Unlike the aspartokinase products of *metL* and *thrA*, the aspartokinase specified by a third locus, *lysC*, is not bifunctional. Nonetheless, the amino acid sequence of this protein exhibits strong homology with the aspartokinase domains, the subunit interaction domains, and the proximal portion of the homoserine dehydrogenase domains of the aspartokinase-homoserine dehydrogenase enzymes I and II. The evidence suggests that the *lysC* locus originated from an ancestral gene that also served as the progenitor of *met* and *thrA* (18). Presumably, in subsequent evolution of the genome, segments required for the homoserine dehydrogenase activity were deleted from the *lysC* locus, retained in *metL* and *thrA*.

Internal Sequence Repetitions

Aspartokinase-homoserine dehydrogenase

The nucleotide sequences specifying some proteins have small internal segments that are repeated within the coding region of the locus. The *thrA* and *metL* genes are not only examples of whole-gene duplications, as discussed above, but they serve as examples of internal duplications as well (34). From the nucleotide sequence of *thrA*, an 820-amino-acid sequence can be inferred which would constitute the primary structure of aspartokinase I-homoserine dehydrogenase I. Likewise, a protein with 809 residues can be derived for aspartokinase II-homoserine dehydrogenase II. When residues 9 to 128 and 146 to 263 of enzyme I are compared, 22% of the nucleotide residues are identical. If residues are scored as matches when no change results in the amino acids specified, the similarity increases to 39%, and if "acceptable" amino acid substitutions are allowed (namely, isoleucine = leucine = valine, serine = threonine, phenylalanine = tyrosine, arginine = lysine, and aspartate = glutamate), the relatedness increases to 44%. These sequences can be viewed as repetitions which occurred within the amino-terminal part of the *thrA* gene and which have diverged over time. A similar analysis of the carboxy end of the gene shows another repetition. Furthermore, one of the repeated segments specifying the carboxy end of the protein has a nucleotide sequence that recurs in the sequence of nucleotides that separates the two large repeats. Thus, the deduced amino acid sequences indicate that both the amino end and the carboxy end of the molecule contain duplicated segments. Analysis of aspartokinase II-homoserine dehydrogenase II reveals that it has a similar organization.

Carbamoylphosphate synthetase

The *carB* locus of *E. coli* specifies a subunit of carbamoylphosphate synthetase. Examination of the 3,219-nucleotide sequence that specifies the sequence of 1,072 amino acids in the *N*-formylmethionine-depleted protein indicates that there is a high level of homology between the amino end and the carboxy end of the enzyme. In particular, when the first 400 residues of the protein are compared with residues 553 to 933 (and adjustments are made for deletions and insertions), 39% of the amino acids are identical. When acceptable amino acid substitutions are scored, the relatedness rises to 64%. Also, when residues 401 to 552 and 933 to 1072 are compared, identities of 20% and relatedness of 45% are seen. Accordingly, the evidence suggests that the coding sequence for this enzyme may also have arisen from internal duplications (85).

Internal repetitions in *thrA*, *thrB*, and *thrC* products

The genes of the *thr* biosynthetic operon share short internal sequences. For the most part, the three genes for biosynthesis of threonine, *thrA*, *thrB*, and *thrC*, are unrelated, as are the bulk of the amino acid residues in the corresponding enzymes. However, there is a 35-amino-acid region that appears twice in the *thrA* enzyme, once in the *thrB* enzyme at a location comparable to one of the *thrA* positions, and once in the *thrC* enzyme at a location comparable to the other position in *thrA*. These sequence similarities could reflect some common functional feature of the three enzymes (90).

Large (>1-kb) Repeated Sequences

E. coli and *S. typhimurium* chromosomes harbor sets of repeated sequences that contribute to rearrangement of the genetic material. Repeated sequences flanking unique sequences have been identified which bear a structural relationship to insertion sequences and transposon organizations and which may bear a functional relationship as well.

rrn loci

Chromosomal rearrangements may come about by unequal crossing over between sister chromatids; the specific sites of crossing over between the sister chromatids can be any repeated sequence. The recurrence of the seven *rrn* loci in *E. coli* and *S. typhimurium* serves as the basis of the formation of a set of duplications and deletions (49, 67, 68), transpositions (51), and inversions (50). Duplications and deletions presumably involve *rrn* loci with the same orientation, inversions involve *rrn* loci with opposite orientations, and transpositions, depending on their character, may involve either orientation.

Hill and collaborators (49) hypothesized that, as a consequence of crossing over between reiterated *rrn* loci, a variety of rare, covalently closed, circular DNAs should be excised from the chromosome of *E. coli* mutants having duplications in the *glyT* region. Such molecules with the predicted contour lengths were isolated. R-loop analyses demonstrated the expected numbers and positions of *rrn* loci in these molecules. These results provided substantive evidence that *rrn* loci served as the basis of deletion of duplicated segments of the genome. The same reasoning indicates that the duplications arose originally through unequal crossing over between *rrn* loci. Similar observations have been made with *S. typhimurium* (67). Furthermore, inversion (50) and transposition (51) of segments of the *E. coli* genome bounded by *rrn* loci also have been demonstrated.

Genetic evidence has been acquired which substantiates that *rrn* loci are associated with duplications. In *S. typhimurium*, detection of duplications at 38 loci has been achieved by first introducing Tn*10* into specific loci (thus creating auxotrophs), then introducing wild-type alleles by transduction, and finally selecting for both prototrophy (a feature of the wild-type allele) and tetracycline resistance (a feature of the Tn*10* insertion in the mutated allele) (3). Mapping the extents of duplications in isolates recovered from such selection regimes revealed duplications bounded by *rrn* loci. Comparisons of recipients transduced to prototrophy alone versus transduction to prototrophy plus antibiotic resistance revealed that duplications are commonplace. Specifically, the segment of the genome bounded by the *rrnB* and *rrnE* loci is merodiploid in approximately 3% of the organisms in a rapidly growing population, whereas other regions bounded by *rrn* loci are duplicated with a frequency between 10^{-4} and 10^{-3} per cell.

rhs loci

E. coli K-12 has at least three *rhs* (rearrangement hot-spot) loci that have highly similar nucleotide sequences, lack a known phenotype, and permit rearrangements reminiscent of those based on *rrn* loci (72). Two of these loci, *rhsA* and *rhsB*, have been mapped and were determined to have the same orientation on the chromosome. The former is approximately 140 kb clockwise from the latter; the sequence of intervening genes is *pit-glyS-xyl*. The *rhs* loci contain 3.8-kb "cores" that can form S1-resistant heteroduplexes. Unequal sister chromatid exchange between *rhsB* and *rhsA* accounts for the observed high frequency of *glyS* duplications. The sizes of *rhs* loci and the frequencies of rearrangements involving these loci are equivalent to the sizes of and rearrangement frequencies involving *rrn* loci. Sequences homologous with *rhs* have not been detected in *S. typhimurium*.

rtl-atl/gat alternation of alleles

The role of flanking sequences in genome organization has been further adduced by studying the relation of some genes that confer the capacity to oxidize polyalcoholic sugars (73). The ability to metabolize ribitol and D-arabitol occurs among 10 to 20% of *E. coli* strains isolated from native habitats. Such metabolic capacity exists in C strains of *E. coli* but not in either B or K-12. The latter strain is able to metabolize galactitol. The genes specifying the ability to use ribitol or D-arabitol are immediately bounded by inverted, imperfectly repeated 1.4-kb sequences. When the operons for pentitol utilization were transduced into strains with functional galactitol genes (*gat*) and selected for pentitol catabolism, the pentitol operons were found to have displaced the *gat* genes. Likewise, when the hexitol genes were transduced into strains with the capacity to metabolize ribitol or D-arabitol, the strains commonly lost the pentitol operons as they gained the *gat* genes (74, 118). Recombination at the flanking sequences is the postulated basis for this allelic relationship. Such reciprocal exclusion is the first example in a monoploid organism of genuine alleles existing as alternative forms having high natural frequencies.

Insertion sequences

Multiple copies of many kinds of insertion sequences are often present in the genomes of *E. coli* and *S. typhimurium* strains, conferring properties of movement and change. Ornithine carbamoyltransferase (as noted earlier) is coded for by the *argI* and *argF* loci; the latter gene is flanked by direct repeats of IS*1*. Selection for hyperproduction of the enzyme can yield mutants which have the *argF* gene amplified as many as 45 times. Examination of restriction endonuclease digests of DNA from mutants indicates that the extent of the amplified sequence corresponds to the distance between the two IS*1* elements. The amplification occurs only when an F factor is integrated nearby in a *cis* position. Whether the reiterated genes are integrated into the chromosome or exist as an extrachromosomal element has not yet been demonstrated (57).

Short (<50-bp) Repeated Sequences

ampC

Direct reiteration of the *ampC* gene of *E. coli* (which specifies β-lactamase) confers resistance to

ampicillin in direct proportion to the number of gene copies. Selection for ever greater levels of resistance allows the isolation of mutants that have DNA segments containing the *ampC* determinant repeated as many as 40 times on the chromosome or 50 times on ColE1-derived plasmids (31). When the nucleotide sequences of the termini of repeated segments, as well as the novel "join point" within the repeated segments, were compared, they were seen to share 12-bp lengths that were identical (32). The sequence showed no elements of symmetry. Analysis of other independently produced duplications revealed that endpoints occurred at different positions in approximately 8 kb of chromosomal DNA. Statistical arguments indicate that, in a random nucleotide sequence, sequences having exactly the same order of 12 nucleotides are likely to occur twice in a length of 4 kb, three times in one of 6.3 kb, and four times in one of 8 kb. The frequency of short nucleotide sequences in the bacterial genome may be modulated such that they contribute to the overall organization of the genome.

lac

A series of compensating mutations at the beginning of the *lac* operon of *E. coli* contribute to a selection scheme that yields amplified segments (112). In this regime, the *lacI* and *lacZ* loci have been fused to produce a hybrid gene lacking the distal portion of the former and the proximal portion of the latter; the gene product, nonetheless, is functional. If expression of the *I* gene is damaged, β-galactosidase production can occur only as a consequence of translation reinitiation sites downstream. When organisms containing all of these genetic alterations are selected for lactose utilization, unstable Lac^+ isolates are obtained. Such mutants contain 40 to 80 copies of the altered *lac* operon on DNA segments ranging in size from 7 to 37 kb (with the representative size ranging from 15 to 20 kb). When the extents of amplified segments in 133 isolates were examined by restriction endonuclease mapping, some preference in the location of endpoints was observed. On the left side of the *lac* operon, the amplified segments originated more or less equally within any of three *Hin*cII fragments and never or rarely originated within two other *Hin*cII fragments. At the right end of the amplified segment, endpoints occurred in a region spanning 14.8 kb. Of these, more than 35% occurred in adjacent 0.4- or 2.6-kb *Hin*cII fragments. Thus, the endpoints of independent recurrences of amplified segments were found to be clustered but not identical.

Short repeated sequences have been observed in the *lacI* region and are known to function in the formation of deletions. The most frequently used sequence (involved in 60% of approximately 250 deletion mutants) was composed of 17 bp. Analysis of the sites at which various deletion mutants originated revealed that 14 of the 17 bp were highly conserved (1). Determining whether such sequences also contributed to the production of *lac* repetitions awaits nucleotide sequencing of the origins, terminations, and join points of the amplified segments.

hin, gin, cin, pin

A site-specific inversion system resides in the *S. typhimurium* chromosome around the H2 antigen locus. The Hin enzyme inverts a segment of about 1 kb of DNA by directing recombination between a pair of flanking inverted repeated sequences of 14 bp each (104). Closely similar site-specific inversion systems exist in phage Mu (*gin*), in phage P1 (*cin*), and within a prophage in the *E. coli* chromosome (*pin*) (22).

REP sequences

Recently, "repetitive extragenic palindromic" (REP) sequences have been identified as major components in the genomes of *E. coli* and *S. typhimurium* (38, 46, 108). In *E. coli*, a consensus sequence of 38 bp was detected in 25% of the operons examined. Extrapolation from this evidence indicates that the REP sequence itself may occur as many as 500 times and may constitute more than 0.5% of the genome. The function of REP sequences has not been established. They appear not to regulate transcription, affect translation, promote gene expression, or function as terminators. However, when inversion mutants in the *his* operon of *S. typhimurium* were selected by having histidine prototrophy restored through fusion of the inverted *hisD* locus to new functional promoters, 7 of 10 inversions originated in an intercistronic REP sequence between *hisG* and *hisD* (100).

hisM

"Odd group" mutants of *S. typhimurium* have completely lost the wild-type capacity to transport L-histidine and have a markedly improved ability to transport L-histidinol. Examination of the nucleotide sequence of the *hisM* gene reveals a 29-bp region that contains multiple repeated sequences; the segment can be viewed as having two direct decanucleotide repeats separated by 2 bp. The nucleotide sequence in seven separately isolated mutants differed from the wild-type sequence in having the 2 bp preceding the decanucleotide repeat, plus the repeat, deleted. These observations establish that a distinct change in substrate specificity can be achieved by a short in-phase deletion of repeated codons (91).

Significance of Repeated Sequences

The recurrent structures that have been described above provide opportunities for restructuring the genetic composition of the bacterial chromosome in which they reside. Enzymes are present that catalyze either general recombination between homologous sequences (e.g., the RecA protein) or sequence-specific recombination (e.g., the Hin enzyme), yet we observe that in nature rearrangements that change the order of genes in *E. coli* or *S. typhimurium* genomes are rare. Much remains unknown about the interactions of the repeated sequences, the features about them that are important, and the mechanisms that regulate the level of the genetic activities of the repeated sequence.

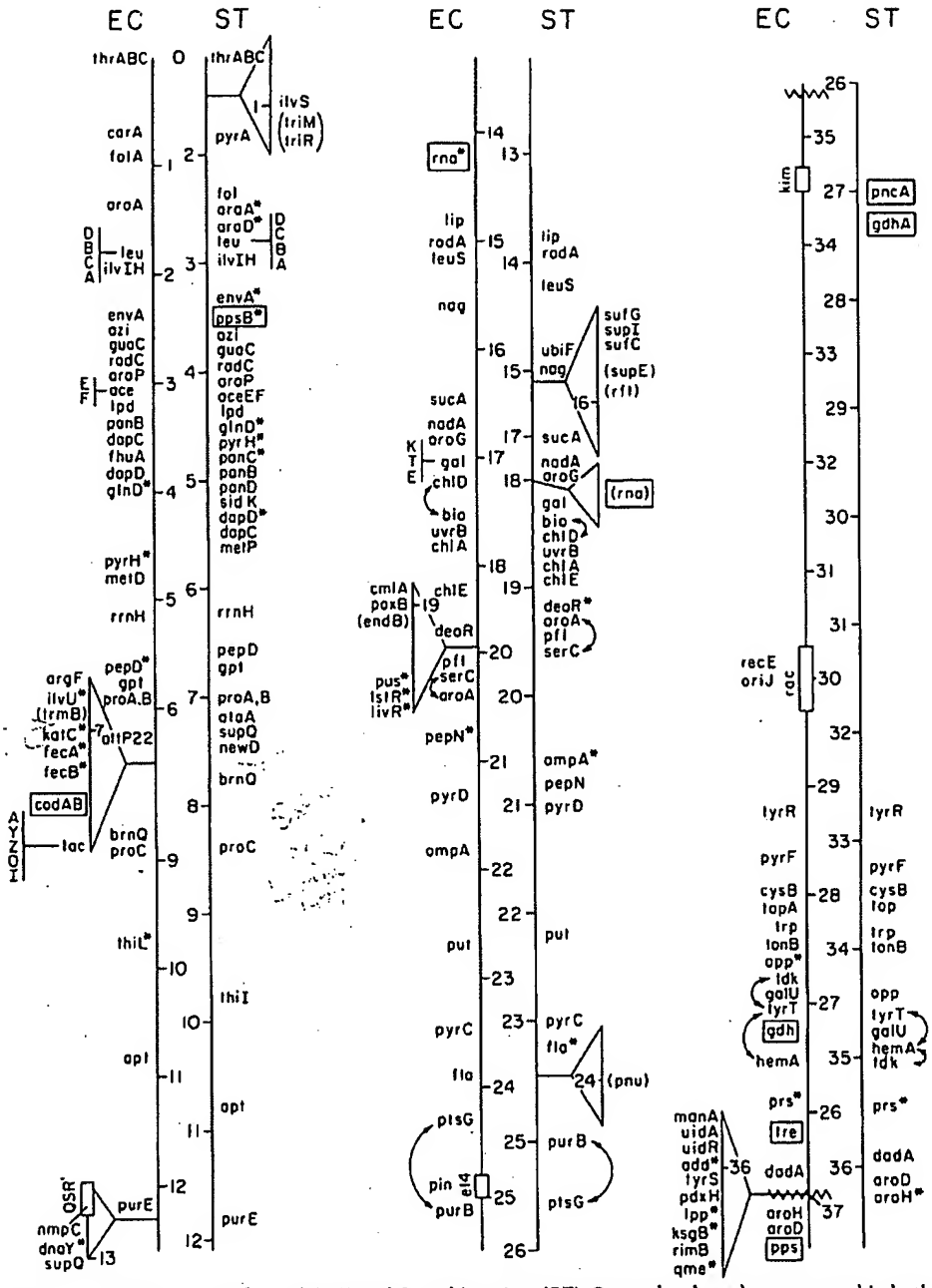

FIG. 1. Alignment of genetic maps of *E. coli* (EC) and *S. typhimurium* (ST). Genes that have been mapped in both *E. coli* (5) and *S typhimurium* (97) were aligned, starting at map position 0. Pairs of genes were not included if one of the two has been mapped only approximately (signified by parentheses in the data sources). In the map alignment, when a pair of genes was displaced by 0.6 map units or more, adjustment was made to improve the alignment. Excess genetic distance in one of the maps is displayed as loops that balloon out from the paired regions to one side or the other. The precise position of the loops in most cases is not closely defined by the available genetic information. Within the constraints of available genetic data, the positions of loops were chosen so as to minimize, where possible, the numbers of known genes on the loops and the total numbers of loops per genome. The zigzag symbols demark the ends of the genetic region that is inverted in one map relative to the other (19). Boxed allele designations indicate genes that reside at markedly different locations in the two maps. Double-headed arrows indicate different map orders for two or more genes in *E. coli* as compared with *S. typhimurium*. (Genes that were marked with an asterisk in the data sources, indicating that they have not been precisely mapped with respect to adjacent markers, were excluded from consideration in this connection.) Rectangular boxes set into the *E. coli* map designate cryptic prophages. The genes for rRNA (the *rrn* series) have been shown to map at analogous locations in *E. coli* and *S. typhimurium* and are omitted from this diagram. All alleles that reside in map segments that are designated as loops are shown in the diagram, including those genes that have been only approximately mapped. Effort has been made to display gene locations accurately; nonetheless, for definitive information, especially in crowded sections, the original compilations of map information should be consulted.

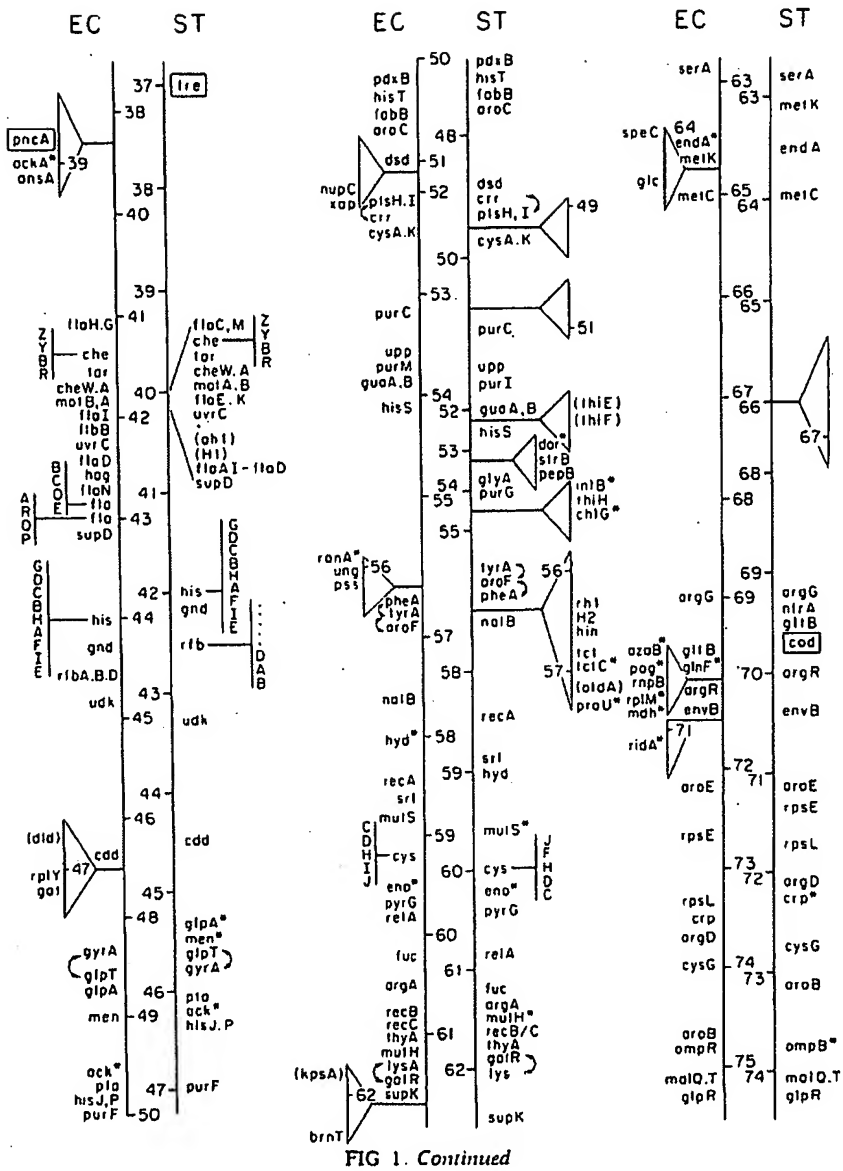

FIG 1. *Continued*

ACCRETIONS TO THE GENOME

Additions and Deletions Suggested by Map Comparison

E. coli and *S. typhimurium* loops

By aligning the genetic maps of *E. coli* and *S. typhimurium*, one sees first of all that the order of genes on the two maps is nearly identical, but also one sees that there are numerous locations where genetic distances between corresponding genes differ. The excess genetic distance in one genome relative to another can be pictured diagrammatically as a bulge or "loop." The two 1983 genetic maps for *E. coli* and *S. typhimurium* were placed side by side and were aligned. Wherever corresponding genes fell greater than 0.6 map units distant from one another, a loop was created to restore alignment. The configuration that resulted is depicted in Fig. 1. Altogether, the *E. coli* map exhibits 14 loops and the *S. typhimurium* map has 15 loops, signifying unequal map distances in5 the two maps, comprising a total of 13.6 map units in the aggregate out of 100 map units total for both organisms. Assuming that map distances are proportional to lengths of DNA, it appears that almost 14% of the DNA of the genomes of both *E. coli* and *S. typhimurium* was acquired by insertion of segments of DNA at the loci of the loops, or lost by deletion of segments of DNA from the partner genome at these same loci, or both. According to this view, the genome of the ancestor of *E. coli* and *S. typhimurium* could have been either larger or smaller than contemporary genomes, according to whether the process of insertion or dele-

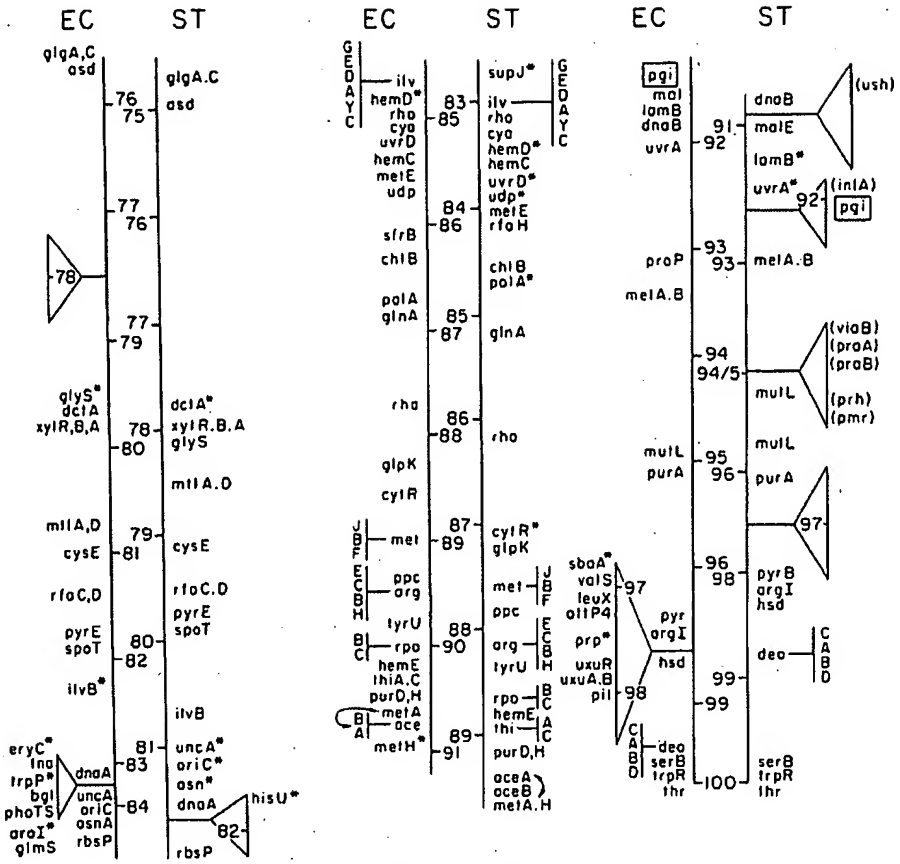

FIG. 1. *Continued*

tion predominated. A schematic representation of the relationship between the two maps is shown in Fig. 2.

Acquisition by transposition

Loops could represent DNA that was acquired from another genetic source and incorporated into the bacterial chromosome. An important mechanism for the acquisition from another genome is transposition. Perhaps at least some of the loops represent acquisition of genetic material horizontally from another genome, events that took place independently in *E. coli* and *S. typhimurium*, involving acquisition of different genetic segments from different sources and a different set of genes in each organism.

Inspection shows that some of the unique phenotypic characteristics that distinguish *E. coli* from *S. typhimurium* are encoded by genes that lie on loops of DNA (Fig. 1), as if these segments of DNA were acquired physically by one genome but not by the other. In *S. typhimurium*, the genes for inositol catabolism, tricarboxylic acid and tricarballylic acid transport, H2 antigen, and the *hin* inversion system are all present in *S. typhimurium*, not in *E. coli*. All of these distinctive genes lie on loops. Similarly, in *E coli*, several distinctive genes, including the supernumerary duplicate *argF* gene, the *lac* operon, and the gene responsible for indole production, *tna*, are all found on loops. Also, the *bgl* and *speC* genes that are present in some *E. coli* strains but not in others are on loops. The genes for galactitol catabolism, the *gat* genes, lie on a loop. As described above, the *gat* genes are subject to replacement in some *E. coli* strains with genes for catabolism of ribitol and D-arabitol.

Some genes appear to have been transposed in the past from one chromosomal location to another in the sense that some genes do not occupy homologous locations but instead occupy different sites in present-day *E. coli* and *S. typhimurium* genomes. In some cases, one of each such a pair of genes is located on a loop. The *cod*, *rna*, *pncA*, and *pgi* genes map differently in the two organisms, and in one genome the gene resides in a loop as if that gene had been relocated as a consequence of transposition of the loop.

Essential genes on loops

Not all loops give the appearance of being recent acquisitions or of being mobile elements at present. Some of the loops contain genes that seem to be

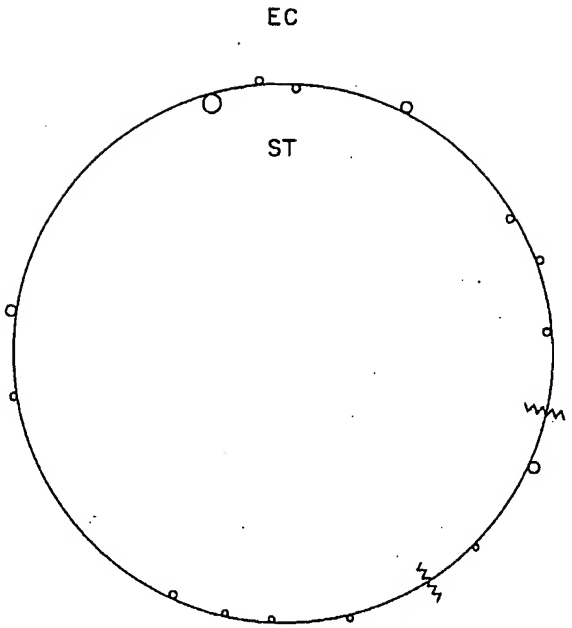

FIG. 2. Schematic representation of the relationship between *E. coli* and *S. typhimurium* genetic maps. The aligned portions of the two maps are represented as a single large circle. Excess map distances are represented as small circles extruded from the large circle, with those for *E. coli* on the outside and those for *S. typhimurium* on the inside of the imaginary heteroduplex.

essential to life and therefore are not the kind of genes one expects to have a history of recent acquisition from another source. Examples are the *E. coli* loop at 36 min that includes the *pdxH*, *lpp*, and *tyrS* genes and the loop at 70 min that includes the *rplY* and *mdh* genes. These are genes for important functions. One would not expect such segments of DNA to have been acquired casually as optional additions to the genome. Other mechanisms must have generated these map position disparities.

Fine structure of loops

According to the arbitrarily chosen rules that were used for the map comparison exercise, the smallest loop structure is 0.6 map units in size. At the upper end, the loops were found to be 1.6 map units in size. If a map unit is taken as 45 kb, the loops range in size from 27 to 72 kb. In terms of individual genetic events, the loops are very large and only coarsely defined. The gross sizes and locations of the structures as illustrated in Fig. 1 may reveal neither the full genetic history nor the complete present organization of any locus, but the diagram does identify genetic regions that may have undergone complex changes in organization during evolutionary divergence. Identification of loop locations can be considered simply as a guide to those regions of the genome that may merit close comparative study. For instance, both ends of the *E. coli* loop at 6.5 map units, the *lac-argF* loop, have been examined in some detail and have provided information on the fine structure of this one loop.

The *argF-lac* loop seems to be about 75 kb long. It is not a single uninterrupted structure, but instead is composed of more than two smaller loops. The *argF* gene, at one end of the loop, is situated within a 12-kb transposonlike segment flanked by two IS*1* elements (54, 122). The *lac* operon, at the other end of the loop, is situated within another 12- to 13-kb loop of DNA which has no detectable homology to *S. typhimurium* DNA and which may have been an active transposon in the past (16, 65). Therefore, the loops of Fig. 1 may appear to be a single, uninterrupted, nonhomologous region, but molecular analysis can reveal complex fine structures which reflect complicated past histories. The true size of DNA segments that have been individually added to or subtracted from the two genomes as a consequence of a single genetic event may be considerably smaller than the size of the loops that are shown in Fig. 1, and the total numbers of addition or deletion events may have been much larger than the numbers of loops in the model.

Additions Suggested by Transposonlike Structures

DNA segments that are too small to be seen in mapping data in terms of the loops may have been acquired by the *E. coli* and *S. typhimurium* genomes. Insertion (IS) elements are not shown on the maps in Fig. 1 both because they are smaller than the arbitrary size chosen for display and because they vary in numbers of copies and genetic locations in different strains of *E. coli* and *S. typhimurium*. The number of copies of IS sequences in a genome is highly variable, not only from species to species but also from strain to strain (54, 110). IS*1*, for instance, appears in one or as many as 30 copies in various *E. coli* strains but is not present in some *S. typhimurium* strains, such as LT2 (83, 84). IS*5* also shows variation in numbers and location in *E. coli* strains (28, 40). The numbers of IS*4* elements vary from strain to strain, oddly enough not independently of the numbers of IS*5* elements. IS*4* differs from the other IS elements in that it seems to reside within a larger transposon, giving it the same flanking sequences regardless of its chromosomal location (28). The *proAB-lac-proC* region of the *E. coli* K-12 genome is highly active in this regard in that it carries a disproportionately large number of IS elements, including IS*1*, IS*5*, IS*21*, and multiple copies of IS*2* and IS*3* (111).

The IS elements in a bacterial chromosome might have been acquired by transposition from infecting plasmids. Not only IS elements, but also functional genes could be acquired from visiting plasmids. As mentioned above, the DNA segment that contains the operons concerned with degradation of ribitol or arabitol is bracketed by large, imperfect repeated sequences, and the supernumerary *argF* gene, a duplicate of *argI*, is flanked by two IS*1* sequences, suggestive in both instances of acquisition by duplicative transposition. The gene for α-hemolysin activity, *hly*, is plasmid borne in some isolates of *E. coli* and is located in the chromosome in other strains. Evidently, the gene has moved from one genome to the other by transposition (20).

Acquisition of Genes from Temperate Phages

Some genes have been acquired as remnants of degenerate forms of temperate prophages. The presence of cryptic prophages in the chromosome is highly variable among *E. coli* strains. Some strains of *E. coli* K-12 carry cryptic lambdoid prophages at map positions 12, 30, and 34 and carry the *e*14 prophage near position 25 min (Fig. 1).

Cryptic lambdoid prophages

At the *rac* locus at map position 30, there are two remnants of a lambdoid phage which contain a recombination function that can be activated by zygotic induction (76) and that can contribute phage recombination genes to defective superinfecting λ *rev* phages (59) The *rac* cryptic prophage can also contribute an alternative replication origin, *oriJ*, to its host chromosome when the bacterial origin of replication, *oriC*, is not functioning (26).

At two separate loci there are similar cryptic lambdoid prophages, one called *qsr'* (sometimes designated QSR') at map position 20 (4, 58) and another called *qin* (sometimes designated *kim*) at map position 34 (11, 33). Both of these defective prophages have the ability to provide phage Q, S, and R functions to superinfecting mutant lambdoid phages (33, 109). The same *qsr'* prophage also contributes a phage gene for outer membrane protein (*nmpC*) to defective mutant bacterial hosts that lack outer membrane protein genes (48) A variant of the *nmpC* gene is also present in phage PA-2 and is expressed in PA-2 lysogens (101) Thus, both cryptic and unimpaired prophages have the capacity to supply phage genes to superinfecting phage mutants and also to supply missing cellular functions to bacterial mutants.

*e*14 prophage

A final example of acquiring genes from phage genomes is the *E. coli pin* gene. The genetic element *e*14 is inducible as a 14.4-kb circle and has a specific attachment site in *E. coli* chromosomes near *purB*, but the defective prophage is present only in some *E. coli* K-12 strains, not in *E. coli* C or B/5 (15, 64). The *e*14 prophage contributes the *pin* gene to the *E. coli* chromosome (92, 93, 113). This locus encodes a specific DNA inversion system that is similar to the *S. typhimurium* Hin system for inverting the H2 antigen segment in *S. typhimurium* and also similar to the related systems in phages P1 and Mu (22). Apparently, the *E. coli* chromosome has acquired the *pin* inversion system by integration of prophage genes.

Clearly some functions in the bacterial genome originated in phage genomes. Ironically, these genetic functions may be "coming home to roost," since phage genes themselves may have originated long ago as fragments of bacterial genomes (30).

SUMMARY

The organizations of the *E. coli* and *S. typhimurium* genomes are highly similar. Gene order has been preserved, with a few exceptions, over a long period of evolutionary time. Both large and small repeated sequences are present in the *E. coli* and *S. typhimurium* genomes at all levels of organization. Intragenomic interactions of some of the repeated sequences can bring about genetic rearrangement. But despite harboring a variety of repeated sequences, the genomes tend not to undergo rearrangements frequently in nature. The mechanisms for regulating and pacing reassortment of bacterial genes have not been identified. Future analysis will, no doubt, lead to the discovery of molecular mechanisms which have determined the organization of the genes and continue to act to maintain the organization of the bacterial genome.

LITERATURE CITED

1 Albertini, A. M., M. Hofer, M. P. Calos, and J. H. Miller. 1982. On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell 29:319–328
2 An, G., and J. D. Friesen. 1980 The nucleotide sequence of *tufB* and four nearby tRNA structural genes of *Escherichia coli* Gene 12:33–39
3. Anderson, P., and J. Roth. 1981 Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (*rrn*) cistrons Proc Natl Acad Sci USA 78:3113–3117.
4 Anilionis, A., P. Ostapchuk, and M. Riley. 1980 Identification of a second cryptic lambdoid prophage locus in the *E coli* K-12 chromosome. Mol. Gen Genet. 180:479–481
5 Bachmann, B. J. 1983. Linkage map of *Escherichia coli* K-12, edition 7 Microbiol Rev. 47:180–230.
6 Bak, A. L., F. T. Black. C. Christiansen, and E. A. Freundt. 1969 Genome size of mycoplasmal DNA Nature (London) 224: 1209–1210.
7. Baptist, J. N., C. R. Shaw, and M. Mandel. 1969 Zone electrophoresis of enzymes in bacterial taxonomy J. Bacteriol. 99:180–188
8 Belfaiza, J., C. Parsot, A. Martel, C. B. de la Tour, D. Margarita, G. Cohen, and I. Saint-Girons. 1986 Evolution in biosynthetic pathways: two enzymes catalyzing consecutive steps in methionine biosynthesis originate from a common ancestor and possess a similar regulatory region Proc Natl Acad. Sci USA 83:867–871
9 Benigni, R., P. A. Petrov, and A. Carere. 1975 Estimate of the genome size by renaturation studies in *Streptomyces*. Appl. Microbiol. 30:324–326.
10 Bollinger, J., C. Park, S. Harayama, and G. L. Hazelbauer. 1984 Structure of the Trg protein: homologies with and differences from other sensory transducers of *Escherichia coli*. Proc. Natl Acad. Sci. USA 81:3287–3291
11 Bouché, J. P., J. P. Galugne, J. Louarn, and L. M. Louarn. 1982 Physical map of a 470 × 10^3 base-pair region flanking the terminus of DNA J Mol. Biol 154:21–32.
12. Boyd, A., K. Kendall, and M. I. Simon. 1983 Structure of the serine chemoreceptors in *Escherichia coli* Nature (London) 301:623–626.
13 Brenner, D. J., G. R. Fanning, F. J. Skerman, and S. Falkow. 1972 Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms J Bacteriol. 109:953–965.
14 Britten, R. J., and D. E. Kohne. 1966. Repeated nucleotide sequences Carnegie Inst Wash. Publ. 66:73–88.
15 Brody, H., A. Greener. and C. W. Hill. 1985 Excision and reintegration of the *Escherichia coli* K-12 chromosomal element e14 J. Bacteriol. 161:1112–1117.
16 Buvinger, W. E., K. A. Lampel, R. J. Bojanowski, and M. Riley. 1984 Location and analysis of nucleotide sequences at one end of a putative *lac* transposon in the *Escherichia coli* chromosome J. Bacteriol. 159:618–623
17 Buxton, R. S., K. Hammer-Jespersen, and P. Valentin-Hansen. 1980. A 2nd purine nucleoside phosphorylase (EC 2.4.2.1) in *Escherichia coli* K-12: xanthosine phosphorylase regulatory mutant isolated as secondary-site revertants of a *deoD* mutant Mol. Gen. Genet 179:331–340
18 Cassan, M., C. Parsot, G. N. Cohen, and J.-C. Patte. 1986 Nucleotide sequence of *lysC* gene encoding the lysine-sensitive aspartokinase III of *Escherichia coli* K12 Evolutionary pathway leading to three isofunctional enzymes J. Biol Chem 261:1052–1057
19 Casse, F., M.-C. Pascal, and M. Chippaux. 1973 Comparisons

between the chromosomal maps of *E. coli* and *S. typhimurium* Length of the inverted segment in the *trp* region Mol Gen Genet. 124:253–257

20 Cavalieri, S. J., G. A. Bohach, and I. S. Snyder. 1984. *Escherichia coli* α-hemolysin: characteristics and probable role in pathogenicity Microbiol Rev 48:326–343.

21 Chow, L. T. 1977 The organization of putative insertion sequences on the *E coli* chromosome, p. 73–79 *In* A I Bukhari. J. A Shapiro. and S L Adhya (ed). DNA insertion elements, plasmids, and episomes Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y

22 Craig, N. L. 1985 Site-specific inversion: enhancers. recombination proteins, and mechanism Cell 41:649–650

23 Crawford, I. P., B P. Nichols, and C. Yanofsky. 1980. Nucleotide sequence of the *trpB* gene in *Escherichia coli* and *Salmonella typhimurium* J Mol Biol 142:489–502.

24 Crosa, J. H., D. J. Brenner, W. H. Ewing, and S. Falkow. 1973. Molecular relationships among the Salmonelleae. J Bacteriol 115:307–315.

25 Davies, W. D., and B. E. Davidson. 1982. The nucleotide sequence of *aroG*, the gene for 3-deoxy-D-arabinoheptulosonate-7-phosphate synthetase (*phe*) in *Escherichia coli* K-12. Nucleic Acids Res. 10:4045–4058

26 Diaz, R., P. Barnsley, and R. H. Pritchard. 1979. Location and characterization of a new replication origin in the *E coli* K-12 chromosome. Mol Gen Genet. 175:151–157

27 Drlica, K. 1984. Biology of bacterial deoxyribonucleic acid topoisomerases. Microbiol Rev 48:273–289.

28 Dykhuizen, D. D., S. A. Sawyer. L. Green. R. D Miller, and D. L. Hartl. 1985 Joint distribution of insertion elements IS*4* and IS*5* in natural isolates of *Escherichia coli* Genetics 111:219–231

29. Dykstra, C. C., D. Prasher, and S. R. Kushner. 1984. Physical and biochemical analysis of the cloned *recB* and *recC* genes of *Escherichia coli* K-12. J. Bacteriol 157:21–27

30. Echols, H. 1979. Bacteriophage and bacteria: friend and foe, p. 487–516 *In* J R Sokatch and L N Ornston (ed.). The bacteria, vol. VII. Mechanisms of adaptation Academic Press. Inc. New York.

31. Edlund, T., T. Grundstrom, and S. Normark. 1979 Isolation and characterization of DNA repetitions carrying chromosomal β-lactamase gene of *Escherichia coli* K-12 Mol. Gen Genet 173:115–125

32 Edlund, T., and S. Normark. 1981 Recombination between short DNA homologies causes tandem duplication Nature (London) 292:269–271.

33 Esplon, D., K. Kaiser, and C. Dambly-Chaudiere. 1983. A third defective lambdoid prophage of *Escherichia coli* K-12 defined by the λ derivative λ*qin*III J Mol. Biol 170:611–633.

34 Ferrara, P., N. Duchange, M. M. Zakin, and G. N. Cohen. 1984 Internal homologies in the two aspartokinase-homoserine dehydrogenases of *Escherichia coli* K-12 Proc Natl Acad. Sci USA 81:3019–3023.

35 Friden, P., J. Donegan, J. Mullen. P. Tsui, M. Freundlich. L. Eoyand, R. Weber, and P. M. Silverman. 1985 The *ilvB* locus of *Escherichia coli* K-12 is an operon encoding both subunits of acetohydroxyacid synthase I. Nucleic Acids Res 13:3979–3993

36 Froshauer, S., and J. Beckwith. 1984. The nucleotide sequence of the gene for *malF* protein, an inner membrane component of the maltose transport system of *Escherichia coli*. J Biol Chem 259:10896–10903

37 Gillis, M., J. De Ley, and M. De Cleene. 1970. The determination of molecular weight of bacterial genome DNA from renaturation rates Eur J. Biochem. 12:143–153.

38. Gilson, M., J.-M. Clement, D. Brutlag, and M. Hofnung. 1984 A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli* EMBO J. 3:1417–1421

39. Goncharoff, P., and B. P. Nichols. 1984 Nucleotide sequence of *Escherichia coli pabB* indicates a common evolutionary origin of *p*-aminobenzoate synthetase and anthranilate synthetase J. Bacteriol 159:57–62

40. Green, L., R. D. Miller, D. E. Dykhuizen, and D. L. Hartl. 1984. Distribution of DNA insertion element *IS5* in natural isolates of *Escherichia coli* Proc Natl. Acad Sci. USA 81:4500–4504

41 Hadley, R. G., M. Hu, M. Timmons. K. Yun, and R. C. Deonier. 1983. A partial restriction map of the PurA-PurE region of the *E. coli* K-12 chromosome. Gene 22:281–287.

42. Harms, E., J.-H. Hsu, C S. Subrahmanyam, and H. E. Umbarger. 1985. Comparison of the regulatory regions of *ilvGEDA* operons from several enteric organisms J Bacteriol 164: 207–216

43 Haughn, G. W., S. R. Wessler, R. M. Gemmil, and J. M. Calvo. 1986 High A+T content conserved in DNA sequences upstream of *leuABCD* in *Escherichia coli* and *Salmonella typhimurium* J Bacteriol. 166:1113–1117.

44. Henson, J. M., and P. L. Kuempel. 1985 Deletion of the terminus region (340 kilobase pairs of DNA) from the chromosome of *Escherichia coli*. Proc. Natl. Acad. Sci. USA 82:3766–3770.

45 Herdman, M. 1985 The evolution of bacterial genomes. p. 37–68. *In* T Cavalier-Smith (ed). The evolution of genome size John Wiley & Sons, Inc., New York.

46 Higgins, C. F., and G. F. L. Ames. 1981 Two periplasmic transport proteins which interact with a common membrane receptor show extensive homology: complete nucleotide sequences Proc Natl. Acad Sci. USA 78:6038–6042

47 Higgins, C. F., G. F. Ames, W. M. Barnes, J.-M. Clement, and M. Hofnung. 1982. A novel intercistronic regulatory element in procaryotic operons Nature (London) 298:760–762

48 Highton, P. J., Y. Chang. W. R. Marcotte. Jr., and C. A. Schnaitman. 1985. Evidence that the outer membrane protein gene *nmpC* of *Escherichia coli* K-12 lies within the defective *qsr'* prophage. J Bacteriol 162:256–262.

49 Hill, C. W., H. Grafstrom, B. W. Harnish, ad B. S. Hillman. 1977 Tandem duplications resulting from recombination between ribosomal RNA genes in *Escherichia coli*. J Mol. Biol 116:407–428.

50 Hill, C. W., and B. W. Harnish. 1981. Inversions between ribosomal RNA genes of *Escherichia coli*. Proc Natl Acad Sci USA 78:7069–7072

51. Hill, C. W., and B. W. Harnish. 1982. Transposition of a chromosomal segment bounded by redundant rRNA genes into other rRNA genes in *Escherichia coli* J. Bacteriol 149:449–457

52 Horowitz, H., G. E. Christie, and T. Platt. 1982 Nucleotide sequence of the *trpD* gene encoding anthranilate synthetase component II of *E coli*. J Mol Biol 156:245–256

53 Houghton, J. E., D. E. Benetne, G. A. O'Donovan, and J. R. Wild. 1984 Protein differentiation: a comparison of aspartate transcarbamoylase and ornithine transcarbamoylase from *Escherichia coli* K-12 Proc. Natl. Acad Sci. USA 81:4864–4868

54 Hu, M., and R. C. Deonier. 1981 Mapping of IS*1* elements flanking the *argF* gene region on the *Escherichia coli* K-12 chromosome Mol Gen Genet. 181:222–229.

55 Hu, M., and R. C. Deonier. 1981 Comparison of insertion sequences IS*1*, IS2, and IS3 copy number in *Escherichia coli* strains K-12. B and C Gene 16:161–170

56 Inokuchi, K., N. Mutoh, S.-I. Matsuyama, and S. Mizushima. 1982. Primary structure of the *ompF* gene that codes for a major outer membrane protein of *Escherichia coli* K-12 Nucleic Acids Res. 10:6957–6968.

57 Jessop, A. P., and C. Clugston. 1985 Amplification of the *Arg*F region in strain HfrP4X of *E coli* K-12. Mol Gen Genet 201: 347–350.

58 Kaiser, K. 1980 The origin of Q-independent derivatives of phage λ. Mol Gen. Genet 179:547–554

59. Kaiser, K., and N. E. Murray. 1979. Physical characterization of the "Rac-prophage" in *E coli* K-12 Mol Gen. Genet 175:159–174

60 Kaplan, J. B., and B. P. Nichols. 1983 Nucleotide sequence of *Escherichia coli pabA* and its evolutionary relationship to *trp*(*G*)*D* J. Mol. Biol 168:451–468

61. Kato, A. C., L. Borstad, M. J. Frazer, and D. J. Denhardt. 1974 Isolation of repeated and self-complementary sequences from *E coli* DNA Nucleic Acids Res 1:1539–1548.

62. Krawiec, S. 1985 Concept of a bacterial species. Int J Syst. Bacteriol 35:217–220

63 Krikos, A., N. Mutoh, A. Boyd, and M. I. Simon. 1983 Sensory transducers of *E. coli* are composed of discrete structural and functional domains. Cell 33:615–622

64. Kutsukake, K., T. Nakao. and T. Iino. 1985. A gene for DNA invertase and an invertible DNA in *Escherichia coli* K-12 Gene 34:343–350

65 Lampel, K. A., and M. Riley. 1982 Discontinuity of homology of *Escherichia coli* and *Salmonella typhimurium* DNA in the *lac* region Mol. Gen. Genet 186:82–86

66. Lehner, A. F., S. Harvey, and C. W. Hill. 1984 Mapping spacer identification of rRNA operons of *Salmonella typhimurium* J Bacteriol. 160:682–686.

67. Lehner, A. F., and C. W. Hill. 1980 Involvement of ribosomal ribonucleic acid operons in *Salmonella typhimurium* chromosomal rearrangements. J Bacteriol. 143:492–498.

68 Lehner, A. F., and C. W. Hill. 1985 Merodiploidy in *Escherichia coli-Salmonella typhimurium* crosses: the role of unequal recombination between ribosomal RNA genes Genetics 110:365–380.

69. Lilley, D. 1986 A new twist to an old story Nature (London) 320:14–15

70 Lilley, D. 1986. Bent molecules—how and why? Nature (London) 320:487–488
71 Lin, H. J. 1974. Isolation of a short, cytosine-rich repeating unit from the DNA of *Escherichia coli* Biochim Biophys Acta 349:13–21
72 Lin, R.-J., M. Capage, and C. W. Hill. 1984 A repetitive sequence, *rhs*, responsible for duplications within the *Escherichia coli* K-12 chromosome J Mol. Biol. 177:1–18
73 Link, C. D., and A. M. Reiner. 1982. Inverted repeats surround the ribitol-arabitol genes of *E coli* C. Nature (London) 298:94–96
74 Link, C. D., and A. M. Reiner. 1983 Genotypic exclusion: a novel relationship between the ribitol-arabitol and galactitol genes of *E coli*. Mol. Gen. Genet. 189:337–339.
75 Louarn, J. M., J. P. Bouché, F. Legendre, J. Louarn, and J. Patte. 1985. Characterization and properties of very large inversions of the *E. coli* chromosome along the origin-to-terminus axis Mol Gen Genet 201:467–476
76 Low, B. K. 1973 Restoration by the *rac* locus of recombinant forming ability in *recB*⁻ and *recC*⁻ merozygotes of *E coli* K-12. Mol Gen Genet 122:119–130.
77 Marmur, J., and P. Doty. 1962 Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. J. Mol Biol 5:109–118.
78 Mizuno, T., M. Y. Chou, and M. Inouye. 1983 A comparative study on the genes for three porins of the *Escherichia coli* outer membrane: DNA sequence of the osmoregulated *ompC* gene J Biol Chem 258:6932–6940.
79 Nakamura, K., and M. Inouye. 1979 DNA sequence of the gene for the outer membrane lipoprotein of *E. coli*: an extremely AT-rich promoter. Cell 18:1109–1117.
80 Nichols, B. P., M. Blumenberg, and C. Yanofsky. 1981 Comparison of the nucleotide sequence of *trpA* and sequences immediately beyond the *trp* operon of *Klebsiella aerogenes*, *Salmonella typhimurium* and *Escherichia coli* Nucleic Acids Res 9:1743–1755
81 Nichols, B. P., G. F. Miozzari, M. van Cleemput, G. N. Bennett, and C. Yanofsky. 1980. Nucleotide sequences of the *trp* regions of *Escherichia coli*, *Shigella dysenteriae*, *Salmonella typhimurium*, and *Serratia marcescens*. J. Mol. Biol 142:503–517.
82 Nichols, B. P., M. van Cleemput, and C. Yanofsky. 1981 Nucleotide sequence of *Escherichia coli trpE* anthranilate synthetase component I contains no tryptophan residues. J Mol. Biol 146:45–54.
83 Nyman, K., K. Nakamura, H. Ohtsubo, and E. Ohtsubo. 1981 Distribution of the insertion element IS*1* in gram-negative bacteria Nature (London) 289:609–612
84 Nyman, K., H. Ohtsubo, D. Davison, and E. Ohtsubo. 1983. Distribution of insertion element IS*1* in natural isolates of *Escherichia coli* Mol. Gen Genet. 189:516–518
85 Nyunoya, H., and C. J. Lusty. 1983. The *carB* gene of *Escherichia coli*: a duplicated gene coding for the carbamoylphosphate synthetase. Proc Natl. Acad. Sci USA 80:4629–4633
86 Ochman, H., and R. K. Selander. 1984 Evidence for clonal population structure in *Escherichia coli* Proc Natl Acad Sci USA 81:198–201.
87 Ochman, H., T. S. Whittam, D. A. Caugant, and R. K. Selander. 1983. Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella* J. Gen Microbiol 129:2715–2726
88 Ogasawara, N., R. Moriya, K. von Meyenberg, F. G. Hansen, and H. Yoshikawa. 1985 Conservation of genes and their organization in the chromosomal replication region of *Bacillus subtilis* and *Escherichia coli*. EMBO J. 4:3345–3350
89 Overbeeke, N., H. Bergmans, F. van Mansfield, and B. Lugtenberg. 1983 Complete nucleotide sequence of *phoE*, the structural gene for the phosphate limitation inducible outer membrane pore protein of *Escherichia coli* K-12 J Mol Biol 163:513–532
90 Parsot, C., P. Cossart, I. Saint-Girons, and G. N. Cohen. 1983. Nucleotide sequence of *thrC* and of the transcription termination region of the threonine operon in *Escherichia coli* K-12. Nucleic Acids Res. 11:7331–7345.
91 Payne, G. M., D. N. Spudich, and G. F.-L. Ames. 1985 A mutational hot-spot in the *hisM* gene of the histidine transport operon in *Salmonella typhimurium* is due to deletion of repeated sequences and results in an altered specificity of transport Mol Gen Genet 200:493–496
92 Plasterk, R. H. A., A. Brinkman, and P. van de Putte. 1983 DNA inversions in the chromosome of *Escherichia coli* and in bacteriophage Mu: relationship to other site-specific recombination systems Proc. Natl Acad Sci USA 80:5355–5358.
93 Plasterk, R. H. A., and P. van de Putte. 1985 The invertible P-DNA segment in the chromosome of *Escherichia coli*. EMBO J 4:237–242.
94 Riley, M., and A. Anilionis. 1978. Evolution of the bacterial genome. Annu. Rev. Microbiol. 32:519–560
95 Rolfe, R., and M. Meselson. 1959 The relative homogeneity of microbial DNA. Proc. Natl Acad Sci USA 45:1039–1043.
96 Sanderson, K. E. 1976. Genetic relatedness in the family *Enterobacteriaceae*. Annu Rev Microbiol. 30:327–349
97. Sanderson, K. E., and J. R. Roth. 1983 Linkage map of *Salmonella typhimurium*, edition VI. Microbiol. Rev. 47:410–453
98. Savić, D. J., S. P. Romac, and S. D. Ehrlich. 1983 Inversion in the lactose region of *Escherichia coli* K-12: inversion termini map within IS3 elements $\alpha_3\beta_3$ and $\beta_5\alpha_5$ J Bacteriol 155:943–946
99 Schildkraut, C., J. Marmur, and P. Doty. 1962. Determination of the base composition of deoxyribonucleic acid from its bouyant density in CsCl. J. Mol. Biol 4:430–443.
100 Schmid, M. B., and J. R. Roth. 1983 Selection and endpoint distribution of bacterial inversion mutations Genetics 105:539–557
101. Schnaitman, C., D. Smith, and M. F. DeSalas. 1975 Temperate bacteriophage which causes the production of a new major outer membrane protein by *Escherichia coli*. J Virol 15:1121–1130.
102 Schultz, J., M. A. Hermodson, G. C. Garner, and K. M. Herrmann. 1984 The nucleotide sequence of the *aroF* gene of *Escherichia coli* and the amino acid sequence of the encoded protein. the tyrosine-sensitive 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase. J. Biol Chem 259:9655–9661
103. Shapiro, H. A. 1970. Distribution of purines and pyrimidines in nucleic acids. p H83–H85. *In* H A. Sober (ed.), Handbook of biochemistry: selected data for molecular biology. 2nd ed CRC Press, West Palm Beach, Fla
104. Simon, M. J., J. Zeig, M. Silverman, G. Mandel, and R. Doolittle. 1980 Phase variation: evolution of a controlling element Science 209:1370–1374.
105 Spencer, M. E., M. G. Darlison, P. E. Stephens, I. K. Duckenfield, and J. R. Guest. 1984 Nucleotide sequence of the *sucB* gene encoding the dihydrolipoamide succinyltransferase of *Escherichia coli* K-12 and homology with the corresponding acetyltransferase Eur. J. Biochem 141:361–374.
106. Squires, C. H., M. DeFelice, J. Devereux, and J. M. Calvo. 1983 Molecular structure of *ilvIH* and its evolutionary relationship to *ilvG* in *Escherichia coli* K12 Nucleic Acids Res. 11:5299–5313
107 Stephens, P. E., M. G. Darlinson, H. M. Lewis, and J. R. Guest. 1983. The pyruvate dehydrogenase complex of *Escherichia coli* K-12: nucleotide sequence encoding the dihydrolipoamide acetyltransferase component Eur J. Biochem. 133:481–489
108 Stern, M. J., G. F. L. Ames, N. H. Smith, E. C. Robinson, and C. F. Higgins. 1984 Repetitive extragenic palindromic sequences: a major component of the bacterial genome Cell 37:1015–1026
109 Strathern, A., and I. Herskowitz. 1975. Defective prophage in *E coli* K-12 strains Virology 41:474–487
110 Timmons, M. S., A. M. Bogardus, and R. C. Deonier. 1983. Mapping of chromosomal IS5 elements that mediate type II F-prime plasmid excision in *Escherichia coli* K-12 J Bacteriol 153:395–407.
111 Timmons, M. S., K. Spear, and R. C. Deonier. 1984. Insertion element IS*J2I* is near *proA* in the chromosomes of *Escherichia coli* K-12 strains. J. Bacteriol 160:1175–1177.
112. Tlsty, T. D., A. M. Albertini, and J. H. Miller. 1984 Gene amplification in the *lac* region of *E coli* Cell 37:217–224.
113 van de Putte, P., R. Plasterk, and A. Kuijpers. 1984 A Mu *gin* complementing function and an invertible DNA region in *Escherichia coli* K-12 are situated on the genetic element e14 J Bacteriol. 158:517–522
114 van Vliet, F., A. Jacobs, J. Plette, D. Gigot, M. Lauwreys, A. Pierard, and N. Glansdorff. 1984 Evolutionary divergence of genes for ornithine and aspartate carbamoyl-transferases—complete sequence and mode of regulation of the *Escherichia coli argF* gene; comparison of *argF* with *argI* and *pyrB* Nucleic Acids Res 12:6277–6289.
115 Wang, E. A., K. L. Mowry, D. O. Clegg, and D. E. Koshland, Jr. 1982 Tandem duplication and multiple functions of a receptor gene in bacterial chemotaxis. J Biol. Chem. 257:4673–4676.
116. Wek, R. C., C. A. Hauser, and G. W. Hatfield. 1985 The nucleotide sequence of the *ilvBN* operon of *Escherichia coli*. sequence homologies of the acetohydroxy acid synthase isozymes Nucleic Acids Res 13:3995–4010
117 Wetmur, J. G., and N. Davidson. 1968 Kinetics of renaturation of DNA J Mol Biol 31:349–370.

118 Woodward, M. J., and H. P. Charles. 1983. Polymorphism in *Escherichia coli: rtl atl* and *gat* regions behave as chromosomal alternatives. J Gen Microbiol. 129:75–84
119. Yanofsky, C. 1984. Comparison of regulatory and structural regions of genes of tryptophan metabolism. Mol Biol. Evol. 1:143–161
120 Yanofsky, C., and M. van Cleemput. 1982 Nucleotide sequence of *trpE* of *Salmonella typhimurium* and its homology with the corresponding sequence of *Escherichia coli* J. Mol Biol 155:235–246
121. Yokata, T., H. Sugisaki, M. Takanami, and Y. Kaziro. 1980 The nucleotide sequence of the cloned *tufA* gene of *Escherichia coli* Gene 12:25–31
122. York, M. K., and M. Stodolsky. 1981. Characterization of P1 *arg* derivatives from *Escherichia coli* K-12 transduction Mol. Gen. Genet. 181:230–240
123 Zengel, J. M., R. H. Archer, and L. Lindahl. 1984 The nucleotide sequence of the *Escherichia coli fus* gene, coding for elongation factor G Nucleic Acids Res 12:2182–2192